

Editorial

Statistical Errors in Clinical Studies

David J. Slutsky, MD¹ Editor-in-Chief¹ The Hand and Wrist Institute, Torrance, California

J Wrist Surg 2013;2:285–287.

Statistics are an integral part of any scientific paper. Unfortunately, statistical errors are common, which can falsely legitimize data. As Song et al¹ noted “The inappropriate use of statistical analysis can lead to incorrect conclusions”. Paul Manske² in his editorial entitled: Lies, Damn Lies and Statistics (quote by British Prime Minister Benjamin Disraeli, 1804–1881) noted that statistical analysis in scientific papers had acquired a shady reputation in part because it had been used to manipulate facts, rather than evaluate them which is embodied by the saying that “Statistics will prove anything, even the truth” (author unknown).

Szabo³ noted that statistics is all about data analysis. Application of the best statistical methodology to poor-quality data is analogous to claiming “the operation was a success

but the patient died.” Knowledge of the appropriate statistical test to apply in any given situation is important, but so is the ability to recognize common statistical errors.

Sample Size

Significance testing is a statistical procedure to determine the probability that the data collected are consistent with the specific hypothesis under investigation. The default position is that there is no relationship between two measured phenomena or that a potential medical treatment has no effect. This is known as the null hypothesis. The investigator's task is to disprove the null to show that a relationship does exist. By convention a P value must be less than 5% ($P < 0.05$) to be statistically significant.

Randomized controlled clinical trials (RCT) that do not show a significant difference between the treatments that are being compared are often called negative. This term wrongly implies that the study has shown that there is no difference, whereas usually all that has been shown is an absence of evidence of a difference. In other words the numbers are too small to make any conclusions.

The sample size of controlled trials is generally inadequate. Many studies are too small to detect even large effects (► **Table 1**). Authors with negative results can therefore not make any meaningful conclusions unless the RCT is sufficiently powered. This error is also frequently found in meta-analyses of published trials, when few or none of the individual trials were statistically large enough. An absence of evidence is not evidence of absence.³

Power Analysis

Typically, the smaller the sample size, the larger any difference between group scores will have to be in order to achieve statistical significance. Statistical power analysis is a set of procedures and formulas that allow us to determine how likely we would achieve statistical significance with a particular sample size given that there is a true difference between groups. If the likelihood is good (e.g. greater than or equal to an 80% chance), then the sample size is considered adequate. A power analysis can be used to calculate the minimum sample size required so that one can be reasonably certain to detect an effect.³ The power of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. As the power increases, the chances of a false negative (Type II error) occurring decreases. In general, the larger the sample the larger the power. Increasing the sample size involves tangible costs, both in time, money, and effort therefore it is important to make the sample size large enough but not wastefully large. The power is typically calculated using a computer software program.

Confidence Interval

A confidence interval (CI) represents the accuracy or precision of a parameter, such as a mean or standard deviation. The CI is a range of values, above and below a finding, in which the actual value is likely to fall. A 95% confidence level, which reflects a significance level of 0.05, means that one would expect 95% of the interval estimates to include the population parameter. The desired level of confidence is set by the researcher and is not determined by the data. A wide confidence interval means that the sample size was too small. A small sample size does not mean that the results are wrong but rather that the data is consistent with a wide range of possible hypotheses. A wide interval cannot provide any meaningful information about the value of a treatment. A narrow or small confidence interval indicates that if we analyzed a different sample we would be reasonably certain that we would get a similar result. A wide confidence interval indicates that we are less sure and perhaps information needs to be collected from a larger number of patients to increase our confidence. Confidence intervals are influenced by the number of people that are being surveyed.

Table 1 After Altman and Bland⁴

Expected difference (P1-P2)	Total sample size required*
5%	1450-3200
10%	440-820
30%	80-100
40%	50-60

5% significant level, 80% power, Small numbers may be justified ($p1 < 0.1$)

Typically, larger surveys will produce estimates with narrow confidence intervals compared to smaller surveys. If outcome measurements are less accurate however it will likely widen the confidence intervals.

Bias

Bias is a general statistical term meaning a systematic (not random) deviation from the true value. A bias of a measurement or a sampling procedure may pose a more serious problem for researcher than random errors because it cannot be reduced by a mere increase in sample size and averaging the outcomes. Randomization is the best way of avoiding bias but it is not always possible or appropriate.

The following examples of bias have been excerpted from a series on Statistics notes in the British Medical Journal (<http://www.users.york.ac.uk/~mb55/pubs/pbstnote.htm>), edited by Doug Altman of Cancer Research UK, and Martin Bland, University of York.⁴

Some biases affecting observational studies:

- Treatment-by-indication bias: different treatments are given to different groups of patients because of differences in their clinical condition.
- Historical controls: will tend to exaggerate treatment effect as recent patients benefit from improvements in health care over time and special attention as a study participant. Recent patients are also likely to be more restrictively selected.
- Retrospective data collection: availability and recording of events and patient characteristics may be related to the groups being compared.

Some biases affecting observational studies and clinical trials:

Selection bias: low response rate or high refusal rate – were patients that participated different to those that did not?

- Informative dropout – was follow-up curtailed for reasons connected to the primary outcome? If so, imbalance in dropout rates between the groups being compared will introduce bias.

Bias in Clinical Trials

No one should know what the next random allocation is going to be as this may affect whether or when the patient is entered into the trial. Using date of birth, hospital number, or simply alter-

nating between treatments is therefore inappropriate. Central randomization is ideal. Unblinded assessment of outcomes may be influenced by knowledge of the treatment group.

Publication Bias

The reviewers for papers submitted for publication do not always agree which papers should be accepted. Because the reviewer's judgments of the quality of papers are therefore made with error, they cannot be perfectly correlated with any measure of the true quality of the paper. In other words, if your manuscript is turned down for journal publication, do not be too despondent. It could be just another example of regression towards the mean.

When reviewing a paper one should look for the following :

- Recognition of the sources of bias and the measures taken to reduce bias through study design
- Selection of patients, collection of data, definition and assessment of outcome and, for clinical trials, method of randomization should be clearly described
- Number and reasons for withdrawal should be reported for the treatment group
- Appropriate analytic methods such as multiple regression should be used to adjust for differences between groups in observational studies
- Authors should discuss likely biases and potential impact on their results

Predictive Values

A common error is to assume that the sensitivity and specificity of a test equates to diagnostic accuracy. The whole point of a diagnostic test is to use it to make a diagnosis, so we need to know the probability that the test will give the correct diagnosis. The sensitivity and specificity do not give us this information. Instead we must approach the data from the direction of the test results, using predictive values. A positive predictive value is the proportion of patients with positive test results who are correctly diagnosed. A negative predictive value is the proportion of patients with negative test results who are correctly diagnosed. If the prevalence of the disease is very low, the positive predictive value will not be close to 1.0 even if both the sensitivity and specificity are high. Thus in screening the general population it is inevitable that many people with positive test results will be false positives.

Matching

In many medical studies a group of cases, people with a disease under investigation, are compared with a group of controls i.e. people who do not have the disease but who are thought to be comparable in other respects. Matching may be by sex, age or ethnic group or other parameters. Matching is done to ensure that the control group and the study subjects are similar in variables to avoid obscuring important differences. Failure to adequately match the study group with the control group can introduce a significant source of error.

This is just a partial list of potential pitfalls to which the author and reviewers must be aware. In this age of evidence based medicine though one must realize however that unlike studying cell cultures, it is not possible to control for the myriad variables that exist in any study involving human subjects. Any given study without statistical significance may still contain some valuable lessons therefore a practical and common sense approach would seem wise.

References

- 1 Song JW, Haas A, Chung KC. Applications of statistical tests in hand surgery. *J Hand Surg* 2009;34(10):1872–1881
- 2 Manske PR. Lies, damn lies and statistics. *J Hand Surg* 1997;22(3):375
- 3 Szabo RM. Statistical analysis as related to hand surgery. *J Hand Surg* 1997;22(3):376–385
- 4 Altman DG, Bland JM. *British Med Journal* 1995;311–485